

S/N 09/559,065

Docket: YOR919990479US1

**IN THE CLAIMS:**

**Please revise the claims, as follows:**

1. (Currently amended) A method for managing and controlling allocation and de-allocation of resources based on a guaranteed amount of resource and additional resources based on a best effort for a plurality of customers, said method comprising:

dynamically allocating server resources for a plurality of customers, such that said resources received by a customer are dynamically controlled and said customer receives a guaranteed minimum amount of resources as specified under a service level agreement (SLA), said service level agreement including at least one parameter defining conditions of dynamically redefining said allocating and de-allocating said server resources; and

designating a service level agreement (SLA) on a server resource for a customer as a form ( $S_{min\#(i)}$ ,  $S_{max\#(i)}$ ,  $M_{bounds(i)}$ ), where  $S_{min\#(i)}$  denotes a guaranteed minimum amount of server resources,  $S_{max\#(i)}$  denotes an upper bound on an amount of server resources that a customer desires to obtain when free resources are available, and  $M_{bounds(i)}$  that includes a low bound ( $M_{lowbound(i)}$ ) and a high bound ( $M_{highbound(i)}$ ) designating bounds on a service level metric for allocating resources beyond the minimum amount  $S_{min\#(i)}$  for each  $i$ -th customer.

2. (Original claim) The method according to claim 1, further comprising:

utilizing a performance metric to increase or decrease an inbound traffic to a customer.

S/N 09/559,065

Docket: YOR919990479US1

3. (Original claim) The method according to claim 1, further comprising:

supporting minimum and maximum server resource-based service level agreements for a plurality of customers.

4. (Original claim) The method according to claim 1, further comprising:

utilizing performance metrics to control the allocation of additional server resources to a plurality of customers using bounds on given service level metrics.

5. (Original claim) The method according to claim 1, further comprising:

supporting a plurality of service level metrics.

6. (Original claim) The method according to claim 1, further comprising:

selectively utilizing a plurality of different metrics for a plurality of different customers.

7. (Original claim) The method according to claim 1, further comprising:

utilizing a service level metric, an amount of allocable resources, and an inbound traffic rate, for defining a state of a current service level ( $M, N, R$ ) for each customer.

8. (Original claim) The method according to claim 1, further comprising:

utilizing a target service level metric  $M_t$  to maintain an actual service level  $M$  substantially at or near a target service level so as to be guaranteed to fall between low and high bounds ( $M_{lowbound}$  and  $M_{highbound}$ ) specified in a service level agreement (SLA).

S/N 09/559,065

Docket: YOR919990479US1

9. (Original claim) The method according to claim 1, further comprising:

computing a target amount of resources  $N_t$  and an inbound traffic rate  $R_t$  from a given target service level metric  $M_t$  and  $(M, N, R)$ .

10. (Original claim) The method according to claim 1, further comprising:

performing at least one of a numerical analysis, a mathematical formulaic operation, an add-one/subtract-one, and a quick simulation for deriving a target amount of resources  $N_t$  and an inbound traffic rate  $R_t$ .

11. (Currently amended) ~~The method according to claim 1, further comprising:~~ A method for managing and controlling allocation and de-allocation of resources based on a guaranteed amount of resource and additional resources based on a best effort for a plurality of customers, said method comprising:

dynamically allocating server resources for a plurality of customers, such that said resources received by a customer are dynamically controlled and said customer receives a guaranteed minimum amount of resources as specified under a service level agreement (SLA), said service level agreement including at least one parameter defining conditions of dynamically redefining said allocating and de-allocating said server resources; and

supporting a resource utilization  $U$  for an actual service level  $M$ , average response time  $T$  for an actual service level  $M$ , and a response time percentile  $T\%$  for an actual service level  $M$ , thereby to support targets of  $U_t$ ,  $T_t$  and  $T_t\%$ .

S/N 09/559,065

Docket: YOR919990479US1

12. (Original claim) The method according to claim 1, further comprising:

deciding whether or not to add a server resource or to reduce an inbound traffic rate to meet service level agreements for a plurality of customers.

13. (Original claim) The method according to claim 1, further comprising:

providing a server farm including means for dynamically allocating servers or server resources to customers as demands of said customers change.

14. (Canceled)

15. (Currently amended) The method according to claim ~~14~~ 1, wherein a minimum amount of server resources  $S_{min\#(i)}$  comprises a guaranteed amount of server resources that the  $i$ -th customer will receive regardless of the server resource usage, and

wherein a maximum amount of server resources  $S_{max\#(i)}$  comprises the upper bound on the amount of server resources that the  $i$ -th customer may receive beyond the minimum amount provided that some unused server resources are available for allocation.

16. (Original claim) The method according to claim 15, wherein a range between  $S_{min\#(i)}$  and  $S_{max\#(i)}$  represents server resources that are provided on an as-available basis, such that the customer is not guaranteed to obtain these resources at any one time, if at all.

17. (Original claim) The method according to claim 1, wherein an allocation of an additional resource is performed so as to keep the performance metric within  $M_{bounds(i)}$ .

S/N 09/559,065

Docket: YOR919990479US1

18. (Original claim) The method according to claim 17, wherein said Mbounds(i) includes any one of bounds on the server resource utilization that are denoted by Ubounds(i), bounds on the average server response time that are denoted by Tbounds(i), and bounds on the server response time percentile that are denoted by T%bounds(i).

19. (Original claim) The method according to claim 1, further comprising:

when a server resource utilization goes above a predetermined set limit

Mhighbound(i), attempting, by a server farm, to maintain the utilization between said predetermined set limits Mbounds(i) by allocating additional server resources to the i-th customer when free resources are available.

20. (Original claim) The method according to claim 19, further comprising:

if free resources are not available, then limiting, by the server farm, an amount of incoming traffic to the i-th customer's server.

21. (Previously presented) The method according to claim 1, further comprising:

controlling said dynamic resource allocation to said plurality of customers to meet a value between minimum and maximum server resources and performance metric-based service level agreements.

S/N 09/559,065

Docket: YOR919990479US1

22. (Original claim) The method according to claim 1, further comprising:

monitoring an inbound traffic rate  $R(i)$ , a currently assigned amount of server resources  $N(i)$ , and a current service level metric  $M(i)$  for all of said plurality of customers.

23. (Original claim) The method according to claim 22, further comprising:

computing a target amount of server resources  $N_t(i)$ , without changing an inbound traffic  $R(i)$ .

24. (Original claim) The method according to claim 23, further comprising:

computing a target inbound traffic rate  $R_t(i)$ , without changing an allocated resource  $N(i)$ , to bring the service level metric  $M(i)$  to the targeted service level metric  $M_t(i)$  from monitored  $R(i)$ ,  $N(i)$  and  $M(i)$  for all  $i$ ,

wherein the target service level metric  $M_t(i)$  comprises the service level metric substantially at or near where  $M(i)$  is to be maintained, and bounded by  $M_{\text{bounds}}(i)$ .

25. (Original claim) The method according to claim 24, further comprising:

determining how to adjust a current  $M(i)$  to the target  $M_t(i)$ , by one of changing  $N(i)$  to  $N_t(i)$  and by bounding the inbound traffic rate  $R(i)$  to  $R_t(i)$ .

26. (Original claim) The method according to claim 25, further comprising:

requesting a system resource manager to perform the resource allocation.

S/N 09/559,065

Docket: YOR919990479US1

27. (Original claim) The method according to claim 26, further comprising:

requesting an inbound traffic controller to throttle an amount of inbound traffic to the plurality of customers.

28. (Original claim) The method according to claim 1, further comprising:

maximizing revenue potential when allocating resources beyond a minimum amount for a customer.

29. (Original claim) The method according to claim 1, wherein a unit of said resources comprises a fixed size unit of allocable or de-allocable resources.

30. (Original claim) The method according to claim 1, wherein a unit of each allocable resource has a different amount.

31. (Previously presented) A method of deciding server resource allocation for a plurality of customers, said method comprising:

computing target values ( $Nt(i)$ ,  $Rt(i)$ ) for every customer  $i$  and setting a variable "TTC-informed( $i$ )" = "no" for all customers " $i$ " such that a record is kept of whether or not throttling on inbound traffic is being applied or not during a given service cycle time;

determining whether or not the service cycle time has expired;

if the service cycle time has not expired, then checking whether an operation state  $M(i)$  is within a predetermined area defined by a metric and a number of resources;

S/N 09/559,065

Docket: YOR919990479US1

if the operation state is not within the predetermined area, then checking whether any customer exists such that a target resource amount  $N_t(i)$  is less than a current resource amount  $N(i)$ ;

if  $N_t(i)$  is less than  $N(i)$ , then determining whether the inbound traffic has been throttled, by determining whether, for any "i",  $ITC\text{-informed}(i) = \text{"yes"}$ ; and

if the inbound traffic has been throttled, then removing the throttling by directing an inbound traffic controller to stop throttling i-th traffic class and setting  $ITC\text{-informed}(i) = \text{"no"}$ .

32. (Original claim) The method according to claim 31, further comprising:

when  $N_t(i)$  is less than  $N(i)$  and it is determined that the inbound traffic is not throttled, deallocating resources from said customers.

33. (Original claim) The method according to claim 32, further comprising:

determining whether the resources must be increased by selecting any i and determining whether  $N_t(i)$  is greater than  $N(i)$ .

34. (Original claim) The method according to claim 33, further comprising:

if it is determined that  $N_t(i)$  is greater than  $N(i)$  and if free resources are judged to be available, then allocating additional resources up to  $N_t(i) - N(i)$  resources without exceeding a maximum amount of server resources  $S_{\max\#(i)}$ .



S/N 09/559,065

Docket: YOR919990479US1

35. (Original claim) The method according to claim 33, further comprising:

if it is determined that  $N_t(i)$  is greater than  $N(i)$  and if free resources are judged to be unavailable and if the currently allocated resource  $N(i)$  is less than the guaranteed minimum  $S_{min\#(i)}$ , then reclaiming resources from those customers  $j$  having more than a guaranteed minimum such that  $N(j) > S_{min\#(j)}$ .

36. (Original claim) The method according to claim 33, further comprising:

if it is determined that  $N_t(i)$  is greater than  $N(i)$  and if free resources are judged to be unavailable and if the currently allocated resource  $N(i)$  is more than or equal to the guaranteed minimum  $S_{min\#(i)}$ , then throttling the inbound traffic.

37. (Original claim) The method according to claim 36, further comprising:

bounding, by the inbound traffic controller, the traffic by  $R_t(i)$ .

38. (Original claim) The method according to claim 31, further comprising:

searching for a potential revenue maximization opportunity when allocating free resources to various customers.

39. (Previously presented) The method according to claim 38, further comprising:

first seeking to de-allocate resources, then allocating additional resources to customers whose service level metric is outside of a predetermined area, and thirdly searching for when the customer's inbound traffic must be throttled due to an exhaustion of free resources or the maximum amount of resources has been already allocated.

S/N 09/559,065

Docket: YOR919990479US1

40. (Currently amended) A system for managing and controlling allocation and de-allocation of resources based on a guaranteed amount of resources and additional resources based on a best effort for a plurality of customers, said system comprising:

a plurality of servers; and

a resource allocation device for dynamically allocating server resources for a plurality of customers, such that said resources received by a customer are dynamically controlled and said customer receives a guaranteed minimum amount of resources as specified under a service level agreement (SLA), said service level agreement including at least one parameter defining conditions of dynamically redefining said allocating and de-allocating said server resources,

wherein a service level agreement (SLA) is designated on a server resource for a customer as a form  $(S_{min\#(i)}, S_{max\#(i)}, M_{bounds(i)})$ , where  $S_{min\#(i)}$  denotes a guaranteed minimum amount of server resources,  $S_{max\#(i)}$  denotes an upper bound on an amount of server resources that a customer desires to obtain when free resources are available, and  $M_{bounds(i)}$  that includes a low bound ( $M_{lowbound(i)}$ ) and a high bound ( $M_{highbound(i)}$ ) designating bounds on a service level metric for allocating resources beyond the minimum amount  $S_{min\#(i)}$  for each  $i$ -th customer.

41-47. (Canceled)

48. (Previously presented) A program product device for storing a program for execution by a digital data processing apparatus to perform a method of deciding server resource allocation for a plurality of customers, said method comprising:

S/N 09/559,065

Docket: YOR919990479US1

computing target values  $(Nt(i), Rt(i))$  for every customer  $i$  and setting a variable  
"ITC-informed(i)" = "no" for all customers "i" such that a record is kept of whether or not  
throttling on inbound traffic is being applied or not during a given service cycle time;  
determining whether or not the service cycle time has expired;  
if the service cycle time has not expired, then checking whether an operation state  
 $M(i)$  is within a predetermined area defined by a metric and a number of resources;  
if the operation state is not within the predetermined area, then checking whether any  
customer exists such that a target resource amount  $Nt(i)$  is less than a current resource  
amount  $N(i)$ ;  
if  $Nt(i)$  is less than  $N(i)$ , then determining whether the inbound traffic has been  
throttled, by determining whether, for any "i",  $ITC\text{-informed}(i) = \text{"yes"}$ ; and  
if the inbound traffic has been throttled, then removing the throttling by directing an  
inbound traffic controller to stop throttling  $i$ -th traffic class and setting  $ITC\text{-informed}(i) =$   
"no".